Student Name: Mai Ngo

Course Name and Number: DSC 575 Intelligent Information Retrieval - SEC 801

Assignment 1-2 - Word Frequency for Index Compression

Date: 1/18/2024

Part B

Question 1: Did the size of vocabulary (2) decrease significantly from [A] to [B]? Why do you think it did/didn't?

Answer 1: The size of vocabulary originally was 17,878 from [A] decreases to 14,714 from [B], a significant 17.698% drop. The noticeable drop happened due to extra word filtering steps: removing stop-words, non-alphabet characters, and lower-case folding. These common characters and words always take a lot of space and have high frequency. Removing these characters/words allow us to narrow content scope and focus on relevant words (high interpretability), ultimately help to distinguish the file and processing volume.

Question 2: Did the size of vocabulary (2) decrease significantly from [B] to [C]? Why do you think it did/didn't?

Answer 2: The size of vocabulary originally was 14,714 from [B] decreases to 10,592 from [C], a significant 28.014% drop. This big drop happened due to stemming application - words are being converted to it root form. For example, the word ‘shares’ is stemmed to ‘share’, thus significantly reduce the number of unique vocabularies, and increase frequency count of the stemmed words.

Question 3: How did the percentage of the top 30 token types (4) change from [B] to [C]? What do you think influenced the change?

Answer 3: The percentage of the top 30 token types originally was 10.453% from [B] increases to 13.571% from [C], a significant 29.829% jump. Certainly, the change happened due to applying Porter stemming. Multiple words that are originally shown as noun, adjective, adverb, plural, past tense, present tense, etc. are all returned as one consolidated stemmed word, thus higher associated frequency count.

In addition, words that are already in top 30 commons in [B] have even higher percentage in [C]. This is because after stemming, those word(s) are now at root form, and able to be combined frequency count with its other original parts of speech form. This ultimately cause the high percentage increment of the top 30 token types.

Reflection: I did not have any hard time completing this assignment. It was very straightforward and easy to figure out. I did challenge myself in re-writing my code more efficiently. For each problem, to get token counts, I started with wring a step by step for-loops (lengthy code), then learn how to rewrite them more efficient and yield shorter running time (ex: stop-words stored as set). I also initially used list accumulator for total number of tokens and keep extending/adding in new tokens lists. More lengthy way but allow me to keep track of my coding logic.

Overall, I think this assignment is reasonable to start with this course, I did not get overwhelmed and can take my time learning the materials. As for jump start the assignment, I already have my own starter code to process the csv file and decided right away to use pandas data frame.

Part A

	[A] Word tokenization (only)	[B] Word tokenization + Case folding (lower-case) + Stop-word filtering + Non-alphabet filtering	[C] Word tokenization + Case folding (lower-case) + Stop-word filtering + Non-alphabet filtering + Porter stemming
(1) Total # of tokens	152000	74131	74131
(2) Size of vocabulary	17878	14714	10592
(3) Top 30 most common token types with frequency (list in descending order of frequency)	[(',', 7382), ('.', 5764), ('the', 5395), ('and', 4264), ('of', 3651), ('to', 3528), ('a', 3505), ('in', 1762), ('--', 1485), ('that', 1472), ("'s", 1217), ('for', 1140), ('``', 898), ("''", 893), ('with', 879), ('we', 878), ('is', 834), ('it', 833), ('?', 824), ('this', 812), ('In', 762), ('on', 761), ('how', 746), ('he', 738), ('from', 707), ('talk', 693), ('his', 686), (':', 643), ('about', 630), ('as', 605)]	[('talk', 700), ('us', 643), ('world', 515), ('new', 415), ('says', 411), ('people', 332), ('shares', 326), ('shows', 282), ('life', 274), ('one', 272), ('ted', 254), ('like', 251), ('make', 239), ('way', 227), ('human', 205), ('work', 203), ('could', 200), ('help', 184), ('even', 179), ('story', 179), ('time', 168), ('years', 163), ('makes', 153), ('talks', 148), ('data', 142), ('future', 142), ('change', 140), ('powerful', 139), ('know', 133), ('two', 130)]	[('talk', 880), ('us', 643), ('world', 527), ('say', 453), ('make', 449), ('share', 444), ('new', 415), ('show', 371), ('use', 360), ('work', 356), ('peopl', 334), ('human', 330), ('way', 326), ('one', 307), ('stori', 307), ('live', 282), ('help', 281), ('life', 274), ('like', 272), ('power', 262), ('ted', 254), ('design', 240), ('take', 223), ('learn', 221), ('look', 219), ('time', 213), ('year', 210), ('think', 204), ('creat', 203), ('could', 200)]
(4) Percentage of tokens in the dataset that is covered by the top 30 token types	35.781%	10.453%	13.571%